SPEECH CODING USING EM SENSOR AND ACOUSTIC SIGNALS

J.F. Holzrichter* and L.C. Ng#
Lawrence Livermore National Laboratory*# and University of California Davis*
P.O. Box 808, L-1, Livermore, CA, 94550
*Person to be contacted: holzrichter1@llnl.gov , 510 658 1128 tele; 510 655 3509 fax

ABSTRACT

Low-power, miniature EM radar-like sensors have made it possible to measure properties of the human speech production system in real-time, without acoustic interference, at low cost. Compression and other applications use an EM sensor measured glottal signal, combined with one or more acoustic signals, to robustly estimate voiced excitation functions, transfer functions, unvoiced speech segments, articulator gestures, and background noise. Applications are speech coding, de-noising, verification, recognition, voice (and music) synthesis, and medical uses. In speech compression, an almost 10-fold bandwidth reduction has been demonstrated, compared to a standard 2.4 kbps LPC10 protocol.

INTRODUCTION

It has been shown that very low power (< 0.3 mW) (EM) radar-like sensors can measure conditions of many of the internal (and external) vocal articulators andvocal tract parameters, in real-time, as speech is generated [1,2,3]. Recent work [4] has determined the details of the physiological processes that lead to EM sensor signals useful for speech processing. These data enable the construction of optimized EM sensors for specific articulator measurements, e.g., vocal folds, vocal tract or oral cavity wall motion, tongue, and other organ movements; and optimized algorithms for specific applications. In particular, a voiced excitation function of speech is obtained by associating EM sensor signals from the glottis with glottal air-flow. These techniques enable accurate transfer function generation, onset of voiced speech, and accurate periods of phonation, robust pitch (<1 Hz accuracy), and, using the statistics of the user's language, enable the definition of periods preceding and following phonation when unvoiced speech is likely to occur.

In addition, they enable the determination of periods of no speech, during which no coding (i.e., no bandwidth) is needed, or during which sampling, processing, and removal of background noise can reliably take place [4] or data can be sent (depending on latency constraints).

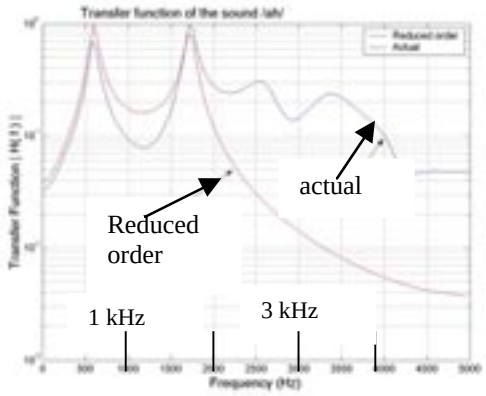

Fig. 1 . Four poles and two zeros per phoneme are needed for acceptable speech intelligibility /ah/

NARROW BANDWIDTH VOCODING

The glottal EM sensor, i.e., GEMs, signal is used in three critical areas: (1) Speech detection and typing, where the speech signal is classified into voiced, unvoiced, and silence segments. (2) The excitation function, from GEMs data, is used to compute a short term transfer function using an autoregressive moving average model. This process yields a pole-zero representation that provides a direct physical mapping to the spectral formants. Since physical vocal articulator motion is slowly varying, the resulting poles and zeros "move" slowly in the complex plane, so the coding algorithm can compact the information into a small number of bits/sec. Experiments have found that a minimum of 4 poles and 2 zeros are needed to model each phoneme as shown in Fig. 1. For compression, this pole-zero model is superior to an

all-pole LPC model, which requires many poles (and bandwidth) to model the presence of zeros in the transfer function. (3) The third advantage is that accurate and economical timing information is generated by using measured pitch periods as basic timing units, and it enables synchronous transfer function processing.

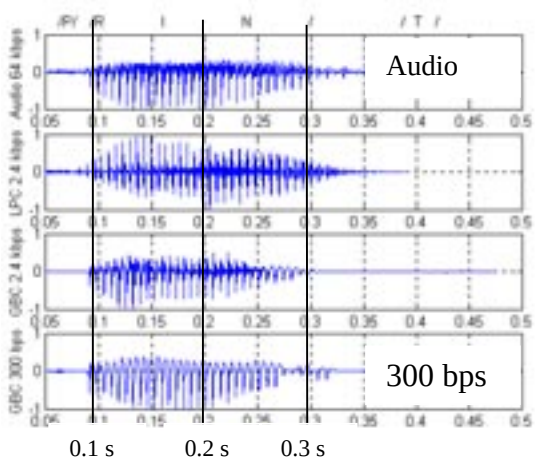

Fig. 2 GBC vocoding of the word “PRINT” at 300 bps

The algorithm first compresses the pole and zero information by utilizing their relatively slow motion on the complex Z plane over a packet transmission interval say one second. Over a ~300ms, the trajectory of poles can be fitted with a cubic polynominal. Thus instead of transmitting 30 coefficients per pole or zero, only 6 coefficients are needed, representing a 5-fold reduction in bandwidth. Similarly, pitch can be coded by a one-time header, with small prosody changes using 3 bps. Unvoiced segments can be represented by a catalog of 8 fricatives using 5 bps, plus amplitude and timing. Thus one arrives at 480 bps for voiced, 280 bps for unvoiced, and 20 bps for silence. Now assuming an occurrence probability of 0.5, 0.2, and 0.3 for each type of speech respectively, the average GBC transmission bandwidth (for this example) becomes approximately 312 bps averaged over a few seconds of speech. Higher bandwidth can be used for improved fidelity

CONCLUSION

The robust measurement of the speech process, made possible with this new class of speech sensors, enables many established as well as new algorithms to be used with confidence and improved effectiveness [5].

ACKNOWLEDGEMENTS

We acknowledge DARPA and NSF SGER for support. Work performed under the auspices of the U.S. DoE by the University of California Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

REFERENCES

[1] Holzrichter, J. F.; Burnett, G. C.; Ng, L. C.; and Lea, W. A., "Speech Articulator Measurements using Low Power EM-wave Sensor," *J. Acoust Soc. Am.* 103 (1) 622, 1998. Also see the Web site http://speech.llnl.gov/ for related information.

[2] Burnett, G. C., "The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract," Thesis UC Davis, ProQuest Digital Dissertations, Inc., Ann Arbor, Michigan, document #9925723, January 15, 1999.

[3] Holzrichter, J.F.; Kobler, J.B.; Rosowski, J.J.; Hillman, R.E.; Ng, L.C.; Burke, G.J.; Champagne II, N.J.; Kallman, J.S.; Sharpe,R.M. “EM Wave Measurements of Glottal Structure Dynamics” UCRL –JC-147775, to be published

[4] Ng, L. C., Burnett, G. C., Holzrichter, J. F. and Gable, T. J. "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing," Lawrence Livermore National Laboratory, UCRL-JC-136631, presented at IEEE ICASSP-2000, Istanbul, Turkey, June 6, 2000.

[5] Aliphcom private communication: see website : www.aliph.com/sound2 for examples.